A novel multi-class SVM model using second-order cone constraints

Julio López[1] · Sebastián Maldonado[2] · Miguel Carrasco[2]

Published online: 9 September 2015

Abstract In this work we present a novel maximum-margin approach for multi-class Support Vector Machines based on second-order cone programming. The proposed method consists of a single optimization model to construct all classification functions, in which the number of second-order cone constraints corresponds to the number of classes. This is a key difference from traditional SVM, where the number of constraints is usually related to the number of training instances. This formulation is extended further to kernel-based classification, while the duality theory provides an interesting geometric interpretation: the method finds an equidistant point between a set of ellipsoids. Experiments on benchmark datasets demonstrate the virtues of our method in terms of predictive performance compared with various other multicategory SVM approaches.

Keywords Multi-class classification · Support vector machines · Second-order cone programming.

✉ Sebastián Maldonado
smaldonado@uandes.cl

Julio López
julio.lopez@udp.cl

Miguel Carrasco
micarrasco@uandes.cl

[1] Facultad de Ingeniería, Universidad Diego Portales, Ejército 441, Santiago, Chile

[2] Facultad de Ingeniería y Ciencias Aplicadas, Universidad de los Andes, Monseñor Álvaro del Portillo 12455, Las Condes, Santiago, Chile

1 Introduction

Multicategory classification is a very important pattern recognition problem in various domains, such as bioinformatics (predicting multiple types of cancer based on microarray data [36]), computer vision applications (classification of different species based on image processing [8]), and business analytic (customer labeling in terms of different levels of risk for a credit institution [9]). A popular machine learning method used to solve this problem is Support Vector Machine (SVM), a convex quadratic programming technique based on the *structural risk minimization* principle, which reduces the risk of *overfitting* and provides better generalization to new data [32].

Although SVM was originally proposed for binary classification, several extensions have been developed to make it suitable for multi-class classification. While most studies in the scientific literature propose splitting the problem into several binary classification problems [14, 27], some approaches attempt to solve a single optimization problem that constructs all classifiers simultaneously [33, 35]. The latter strategy takes all available information into account, which may lead to superior predictive performance, especially in low dimensional datasets.

Second-order cone programming SVM (SOCP-SVM) [26] is a recently proposed alternative for classification purposes. This method constructs a maximum-margin classifier in such a way that the false positive and false negative error rates do not exceed a predefined value [26], and it is based on a robust setting for class-conditional densities. SOCP formulations are special cases of nonlinear convex optimization problems, which can be solved via interior point algorithms [4].

The second-order cone programming SVM formulation proposed by Nath and Bhattacharyya [26] has been applied successfully for binary classification [21], and we propose an extension to multicategory learning in this work. Our proposal constructs the classification hyperplanes "all-together" in a single SOCP formulation, providing a robust and powerful method for performing this task. Our method has important differences compared with the work proposed by Zhong and Fukushima [37], in which the chance constraints are designed to deal with noisy data (instances with measurement errors for example), instead of providing a probabilistic framework for the class-conditional densities. Our approach also differs from Debnath et al. [11] strategy, where the authors solve the standard SVM formulation for binary classification via a SOCP optimization scheme.

This paper has the following structure: Section 2 introduces SVM for multicategory classification. The proposed SOCP-SVM approach is presented in Section 3 in its linear version, while Section 4 extends these ideas to kernel functions. In Section 5 we propose multi-class extensions based on the work by Nath and Bhattacharyya [26] for benchmarking purposes. Section 6 provides the experimental results using benchmark datasets. A summary of this work can be found in Section 7, where we provide its main conclusions and address future developments.

2 Prior work in support vector classification

The most commonly used multi-class SVM formulations (OvO-SVM, OvA-SVM and MC-SVM), which will be used as alternative methods in our experiments, are described in this section. Additionally, we present two important "all-together" SVM methods which are closely related to our proposal: Scatter-SVM [16] and the multicategory SVM formulation proposed by Bredensteiner and Bennet [10], and by Yajima [35].

2.1 One-versus-all approach

This is the simplest and probably the earliest alternative for multicategory SVM [32]. For m training samples of the form $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$, where $\mathbf{x}_i \in \Re^n$ is the i-th instance and $y_i \in \{1, 2, \ldots, K\}$ its respective class label, this method constructs K binary SVM classifiers, each one of which aims at separating one category from the remaining classes. The k-th model of OVA-SVM has the following form:

$$\begin{array}{ll} \min\limits_{\mathbf{w}_k, b_k, \boldsymbol{\xi}_k} & \frac{1}{2}\|\mathbf{w}_k\|^2 + C\sum\limits_{i=1}^{m}\xi_i^k \\ \text{s.t.} & \tilde{y}_i\left(\mathbf{w}_k^\top \cdot \mathbf{x}_i + b_k\right) \geq 1 - \xi_i^k, \\ & \qquad \xi_i^k \geq 0, \quad i = 1, \ldots, m, \end{array} \tag{1}$$

where $\tilde{y}_i = 1$ means the sample i has label k ($y_i = k$), while $\tilde{y}_i = -1$ corresponds to the opposite case: object i belongs to a different category from k's. The decision function for OVA-SVM is given by $f^k(\mathbf{x}) = \mathbf{w}_k^\top \cdot \mathbf{x} + b_k$, and a new sample $\mathbf{x}$ is assigned to the class with the greatest value of $f^k(\mathbf{x})$ (i.e. $f^{k^*}(\mathbf{x}) = \max\{f^k(\mathbf{x}) : k = 1, \ldots, K\}$).

2.2 One-versus-One approach

Another well-known SVM variation is known as One-versus-One (OvO) SVM [17], which constructs $K(K-1)/2$ binary SVM classifiers, one for each pair of categories. Given training points from classes k and l, OvO SVM solves the following problem:

$$\begin{array}{ll} \min\limits_{\mathbf{w}_{kl}, b_{kl}, \boldsymbol{\xi}^{kl}} & \frac{1}{2}\|\mathbf{w}_{kl}\|^2 + C\sum\limits_{r}\xi_r^{kl} \\ \text{s.t.} & \mathbf{w}_{kl}^\top \cdot \mathbf{x}_r + b_{kl} \geq 1 - \xi_r^{kl}, \text{ if } y_r = k, \\ & -(\mathbf{w}_{kl}^\top \cdot \mathbf{x}_r + b_{kl}) \geq 1 - \xi_r^{kl}, \text{ if } y_r = l, \\ & \qquad \xi_r^{kl} \geq 0, \quad r = 1, \ldots, m_k + m_l, \end{array} \tag{2}$$

where m_k and m_l are the cardinality of the sets of training points of classes k and l, respectively. The decision function for a new instance $\mathbf{x}$ is given by $f^{kl}(\mathbf{x}) = \mathbf{w}_{kl}^\top \cdot \mathbf{x} + b_{kl}$. A Max-Wins voting strategy is used, in which each classification function assigns its respective data objects to one of the two categories, increasing the vote for the assigned class by one [12]. The category with most votes determines the classification of each new object.

2.3 "All-together" SVM approaches

Several multi-class SVM approaches that solve one single optimization problem have been proposed in the literature. The MC-SVM method [33] extends the ideas of OVA-SVM by constructing K binary classifiers simultaneously. The formulation of this approach follows:

$$\begin{array}{ll} \min\limits_{\tilde{\mathbf{w}}, \mathbf{b}, \boldsymbol{\xi}} & \frac{1}{2}\sum\limits_{k=1}^{K}\|\mathbf{w}_k\|^2 + C\sum\limits_{i=1}^{n}\sum\limits_{k=1, k\neq y_i}^{K}\xi_i^k \\ \text{s.t.} & (\mathbf{w}_{y_i}^\top \cdot \mathbf{x}_i + b_{y_i}) - (\mathbf{w}_k^\top \cdot \mathbf{x}_i + b_k) \geq 2 - \xi_i^k, \\ & \quad \xi_i^k \geq 0, \quad i = 1, \ldots, m, \; k \in \{1, \ldots, K\} \setminus y_i, \end{array} \tag{3}$$

where $\tilde{\mathbf{w}} = [\mathbf{w}_1^\top, \mathbf{w}_2^\top, \ldots, \mathbf{w}_K^\top]^\top \in \Re^{nK}$ and $\mathbf{b} = [b_1, b_2, \ldots, b_K]^\top \in \Re^K$ represent all the hyperplanes constructed by this approach. The decision rule is similar to that of the OvA SVM formulation, where a new sample $\mathbf{x}$ belongs to the class k^* iff $k^* = \text{argmax}_{k=1,\ldots,K}\{\mathbf{w}_k^\top \cdot \mathbf{x} + b_k\}$. An alternative formulation inspired in OVO-SVM can be found in Yajima [35]. In that paper, the author proposes the following quadratic problem, which subsequently transforms to a linear programming formulation via the l_1 regularization:

$$\begin{aligned} \min_{\widetilde{\mathbf{w}},\mathbf{b},\boldsymbol{\xi}} \quad & \frac{1}{2}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} \|\mathbf{w}_k - \mathbf{w}_l\|^2 + \frac{\theta}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|^2 + C\sum_{k=1}^{K}\sum_{l=1,l\neq k}^{K} \mathbf{e}^\top \cdot \boldsymbol{\xi}^{kl} \\ \text{s.t.} \quad & (\mathbf{w}_k^\top \cdot \mathbf{x}_i + b_k) - (\mathbf{w}_l^\top \cdot \mathbf{x}_i + b_l) \geq 1 - \xi_i^{kl}, \\ & \xi_i^{kl} \geq 0, \quad i = 1, \ldots, m_k;\ k, l \in \{1, \ldots, K\}\, k \neq l, \end{aligned} \tag{4}$$

where $\boldsymbol{\xi}^{kl} \in \Re^{m_k}$ and $\mathbf{e}$ denote an all-ones vector of appropriate dimension.

The parameter θ controls the trade-off between the errors and the regularization term. This parameter is related to the Tikhonov regularization [31] and the viscosity methods [6] and is usually used to avoid ill-conditioning problems. The case of $\theta = 1$ was first studied by Bredensteiner and Bennet [10].

Another multi-class approach was proposed by Jensen et al. [16] for ν-SVM (Scatter-SVM) and Ñanculet et al. [1] for standard SVM (AD-SVM formulation). The main idea is to find a point that is equidistant to all classes that minimizes the distance between their reduced convex hulls (also referred to the *center of the configuration* [1]). This can be obtained by solving the following minimization problem, which results from the dual formulation of AD-SVM:

$$\begin{aligned} \min_{\tilde{\mathbf{w}},\mathbf{b},\boldsymbol{\xi}} \quad & \frac{1}{2}\sum_{k=1}^{K}\|\mathbf{w}_k - \bar{\mathbf{w}}\|^2 + C\sum_{k=1}^{K}\mathbf{e}^\top \cdot \boldsymbol{\xi}_k \\ \text{s.t.} \quad & X_k^\top(\mathbf{w}_k - \bar{\mathbf{w}}) + b_k\mathbf{e} + \boldsymbol{\xi}_k \geq \mathbf{e}, \quad k = 1, \ldots, K, \\ & \bar{\mathbf{w}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k, \quad \sum_{k=1}^{K} b_k = 0, \end{aligned} \tag{5}$$

where $\boldsymbol{\xi} = [\boldsymbol{\xi}_1^\top, \ldots, \boldsymbol{\xi}_K^\top]^\top \in \Re^m$, X_k is the matrix of all training points of class k, and $\bar{\mathbf{w}}$ represents the center of the configuration.

Alternatively to discriminative methods, some "all-together" SVM-based approaches that follow a different classification strategy have been recently proposed. One of such methods is MSM-SVM, a maximal-margin spherical-structured multi-class approach based on the principles of one-class classification for outlier detection [13]. This technique constructs hyperspheres that tightly enclosed each training pattern while controlling the number of support vectors. MSM-SVM is potentially useful in multi-class problems with skewed class distributions.

3 Proposed linear SOCP-SVM formulation

In this section, we present a novel multi-class linear SVM formulation using second-order cones, for which all classifiers are constructed simultaneously. The reasoning behind this approach is that we can construct the classifiers by finding a new center of the configuration, which would be a point, equidistant to all classes, that minimizes the distance between the ellipsoids which represent each class, instead of the reduced convex hulls. In this section we describe the notation and preliminaries regarding second-order cone programming first. Next, the primal form of the proposed approach is presented, while the geometrical interpretation based on the dual form of the approach is discussed at the end of this section. The kernel version of our approach is presented in Section 4.

3.1 Notation and preliminaries

Let us denote the set of points $\mathbf{x}_i$ such that $y_i = k$ by $\mathcal{A}^k$, and by m_k its cardinality. Let $\mathbf{X}_k$ be a random vector variable that generates the sample $\mathcal{A}^k$, with mean $\boldsymbol{\mu}_k \in \Re^n$ and covariance matrix Σ_k for $k = 1, \ldots, K$, where $\Sigma_k \in \Re^{n\times n}$ are symmetric positive semi-definite matrices. Let us denote a family of distributions which have a common mean and covariance by $\mathbf{X} \sim (\boldsymbol{\mu}, \Sigma)$. For binary classification, Nath and Bhattacharyya [26] proposed the following probabilistic constraints:

$$\Pr\left\{\mathbf{w}^\top\mathbf{X}_1 + b \geq 1\right\} \geq \eta_1, \quad \Pr\left\{\mathbf{w}^\top\mathbf{X}_2 + b \leq -1\right\} \geq \eta_2. \tag{6}$$

The above constraints suggest that the probability of false-negative and false-positive errors should not exceed a predefined parameter $1 - \eta_k$ with $\eta_k \in (0, 1]$, for each class k. Although this has been suggested in the literature, Nath and Bhattacharyya showed evidence that test errors may exceed these thresholds, and suggest setting these parameters via cross validation since their interpretation is not straightforward [26].

In order to classify each training pattern k correctly up to the rate η_k, even for the worst data distribution, the probabilistic constraints are then replaced with their robust counterparts:

$$\begin{aligned} \inf_{\mathbf{X}_1\sim(\boldsymbol{\mu}_1,\Sigma_1)} \Pr\left\{\mathbf{w}^\top\mathbf{X}_1 + b \geq 1\right\} \geq \eta_1, \\ \inf_{\mathbf{X}_2\sim(\boldsymbol{\mu}_2,\Sigma_2)} \Pr\left\{\mathbf{w}^\top\mathbf{X}_2 + b \leq -1\right\} \geq \eta_2. \end{aligned} \tag{7}$$

The previous constraints can be converted into second-order cones thanks to the application of the Chebyshev inequality [18, Lemma 1]. Taking this into account, the SOCP-SVM formulation for binary classification follows [26]:

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} \quad & \mathbf{w}^\top\boldsymbol{\mu}_1 + b \geq 1 + \kappa_1\sqrt{\mathbf{w}^\top\Sigma_1\mathbf{w}}, \\ & -(\mathbf{w}^\top\boldsymbol{\mu}_2 + b) \geq 1 + \kappa_2\sqrt{\mathbf{w}^\top\Sigma_2\mathbf{w}}, \end{aligned} \tag{8}$$

where $\kappa_i = \sqrt{\frac{\eta_i}{1-\eta_i}}$, for $i = 1, 2$. The parameter κ_i controls the size of the ellipsoids [18], and the problem may become infeasible for some η_i due to the intersection of ellipsoids [26]. This can be avoided by introducing slack variables, as was proposed by Maldonado and López in [21].

Next, we describe our proposal in its primal formulation based on SOCP-SVM and the concept of the center of the configuration.

3.2 Primal formulation for multi-class SOCP-SVM

Motivated by the studies described in Section 2.3, we suggest considering the following quadratic chance-constrained programming problem with an additional regularization term:

$$
\begin{aligned}
\min_{\tilde{\mathbf{w}},\mathbf{b}} \quad & \frac{1}{2}\sum_{k=1}^{K-1}\sum_{l=k+1}^{K} \|\mathbf{w}_k - \mathbf{w}_l\|^2 + \frac{\theta}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|^2 \\
\text{s.t.} \quad & \Pr\left\{(\mathbf{w}_k - \bar{\mathbf{w}})^\top \cdot \mathbf{X}_k + b_k - 1 \geq 0\right\} \geq \eta_k, \quad k = 1, \ldots, K, \\
& \bar{\mathbf{w}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k, \quad \sum_{k=1}^{K} b_k = 0,
\end{aligned}
\tag{9}
$$

where $\tilde{\mathbf{w}} = \left[\mathbf{w}_1^\top, \mathbf{w}_2^\top, \ldots, \mathbf{w}_K^\top\right]^\top \in \Re^{nK}$, $\mathbf{b} = [b_1, b_2, \ldots, b_K]^\top \in \Re^K$, $\eta_k \in (0, 1)$ and $\theta \geq 0$ denotes the control parameter. It is important to notice that $\sum_{k=1}^{K-1}\sum_{l=k+1}^{K}\|\mathbf{w}_k - \mathbf{w}_l\|^2$ is equivalent to $\sum_{k=1}^{K}\|\mathbf{w}_k - \bar{\mathbf{w}}\|^2$, i.e. the differences between each pair of weight vectors lead to the concept of center of the configuration. Following the procedure presented in Nath and Bhattacharyya [26], Formulation (9) becomes:

$$
\begin{aligned}
\min_{\tilde{\mathbf{w}},\mathbf{b}} \quad & \frac{1}{2}\sum_{k=1}^{K}\|\mathbf{w}_k - \bar{\mathbf{w}}\|^2 + \frac{\theta}{2}\sum_{k=1}^{K}\|\mathbf{w}_k\|^2 \\
\text{s.t.} \quad & \kappa_k \left\|S_k^\top(\mathbf{w}_k - \bar{\mathbf{w}})\right\| \leq (\mathbf{w}_k - \bar{\mathbf{w}})^\top \boldsymbol{\mu}_k + b_k - 1, \ k = 1, \ldots, K, \\
& \bar{\mathbf{w}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_k, \quad \sum_{k=1}^{K} b_k = 0,
\end{aligned}
\tag{10}
$$

where $\Sigma_k = S_k S_k^\top$ and $\kappa_k = \sqrt{\frac{\eta_k}{1-\eta_k}}$, for $k = 1, \ldots, K$.

Next, following the notation presented in [35], we rewrite Formulation (10) in a compact form. Let us denote by

$$
\mathbf{Q}(\theta) = (K + \theta)I_{nK} - \mathcal{J} \in \Re^{nK \times nK}, \tag{11}
$$

with

$$
\mathcal{J} = \begin{bmatrix} I_n & I_n & \cdots & I_n \\ I_n & I_n & \cdots & I_n \\ \vdots & \vdots & \ddots & \vdots \\ I_n & I_n & \cdots & I_n \end{bmatrix} \in \Re^{nK \times nK}.
$$

where I_{nK} and I_n denote the identity matrix of size nK and n, respectively.

Note that the matrix $\mathbf{Q}(0)$ is symmetric positive semi-definite, and that the matrix$\mathbf{Q}(\theta)$ is symmetric positive definite for $\theta > 0$ (see [35, Proposition 3.3, Proposition 3.4]). Then, the objective function of problem (10) can be expressed as:

$$
\frac{1}{2}\tilde{\mathbf{w}}^\top \mathbf{Q}(\theta)\tilde{\mathbf{w}} = \frac{1}{2}\|\mathbf{Q}^{1/2}(\theta)\tilde{\mathbf{w}}\|^2, \tag{12}
$$

where

$$
\mathbf{Q}^{1/2}(\theta) = \sqrt{K + \theta}\, I_{nK} - \frac{\sqrt{K + \theta} - \sqrt{\theta}}{K}\mathcal{J}. \tag{13}
$$

Let H^i be the $n \times nK$ matrix with all blocks being $-\frac{1}{K}I_n$ except the ith block being $(1 - \frac{1}{K})I_n$, that is,

$$
H^i = \left[-\frac{1}{K}I_n, \ldots, -\frac{1}{K}I_n, \left(1 - \frac{1}{K}\right)I_n, -\frac{1}{K}I_n, \ldots, -\frac{1}{K}I_n\right], \ i = 1 \ldots, K.
$$

Then,

$$
\mathbf{w}_i - \bar{\mathbf{w}} = H^i\tilde{\mathbf{w}}. \tag{14}
$$

Let us denote the K-dimensional canonical vector by $\mathbf{d}^i$, that is,

$$
\mathbf{d}^i = [0, \ldots, 0, 1, 0, \ldots, 0]^\top.
$$

With this,

$$
b_i = (\mathbf{d}^i)^\top \mathbf{b}. \tag{15}
$$

Then, by using previous definition (12)–(15), Problem (10) can be rewritten compactly as follows:

$$
\begin{aligned}
\min_{\tilde{\mathbf{w}},\mathbf{b}} \quad & \tfrac{1}{2}\|\mathbf{Q}^{1/2}(\theta)\tilde{\mathbf{w}}\|^2 \\
\text{s.t.} \quad & \kappa_i\|S_i^\top H^i\tilde{\mathbf{w}}\| \leq (H^i\tilde{\mathbf{w}})^\top\boldsymbol{\mu}_i + (\mathbf{d}^i)^\top\mathbf{b} - 1, \ i = 1, \ldots, K, \\
& \mathbf{e}^\top\mathbf{b} = 0,
\end{aligned}
\tag{P_θ}
$$

where $\mathbf{e}$ denotes a vector of ones of dimension K.

By introducing a new variable t, Formulation (P_θ) can be written equivalently as the following problem:

$$
\begin{aligned}
\min_{t,\tilde{\mathbf{w}},\mathbf{b}} \quad & t \\
\text{s.t.} \quad & \begin{bmatrix} t - 1 \\ \sqrt{2}\mathbf{Q}^{1/2}(\theta)\tilde{\mathbf{w}} \end{bmatrix} \leq t + 1, \\
& \kappa_i\|S_i^\top H^i\tilde{\mathbf{w}}\| \leq (H^i\tilde{\mathbf{w}})^\top\boldsymbol{\mu}_i + (\mathbf{d}^i)^\top\mathbf{b} - 1, \ i = 1, \ldots, K, \\
& \mathbf{e}^\top\mathbf{b} = 0,
\end{aligned}
\tag{16}
$$

which is a convex programming problem with a linear objective function, $K + 1$ second-order cone (SOC) constraints, and one affine equality constraint. We recall that an SOC constraint (see [3] for details) on the variable $\mathbf{x} \in \Re^n$ can be expressed as:

$$
\|A\mathbf{x} + \mathbf{b}\| \leq \mathbf{c}^\top\mathbf{x} + d,
$$

where $d \in \Re$, $\mathbf{c} \in \Re^n$, $\mathbf{b} \in \Re^m$, $A \in \Re^{m \times n}$ are given. Thus, Problem (P_θ) can be cast as a linear second-order cone programming (SOCP) problem.

Problem (P_θ) will be called (linear) Center-of-the-Configuration SOCP-SVM formulation (CC-SOCP-SVM$_l$). The solution of this problem leads to the construction of K classifiers, and a new point $\mathbf{x}$ will be classified as the class which attains the greatest value of $f^k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + b_k$.

3.3 Dual formulation and geometric interpretation

In this section we present first the dual formulation of the Multiclass SOCP-SVM, which provides an interesting geometric interpretation for this problem. Additionally, we discuss the relationship between our approach and others reported in the literature.

The following theorem gives the dual formulation of problem (P_θ) where $\theta \geq 0$. This formulation is interpreted geometrically as the minimization of distances between K ellipsoids (see Proposition 1 below).

Theorem 1 *For a given $\theta \geq 0$, the dual problem of Problem (P_θ) is given by:*

$$\begin{aligned} \max_{z_i, \boldsymbol{u}^i, \beta} \quad & K\beta - \tfrac{\beta^2}{2(K+\theta)} \sum_{i=1}^{K} \|z_i - \boldsymbol{p}\|^2 \\ \text{s.t.} \quad & z_i = \boldsymbol{\mu}_i - \kappa_i S_i \boldsymbol{u}^i, \ i = 1, \ldots, K, \\ & \|\boldsymbol{u}^i\| \leq 1, \ i = 1, \ldots, K, \\ & \boldsymbol{p} = \tfrac{1}{K} \sum_{i=1}^{K} z_i, \\ & \beta \geq 0. \end{aligned} \tag{17}$$

The proof of Theorem 1 is presented in the Appendix A.1. The following proposition, derived from problem (17), gives us a geometric interpretation of the formulation (P_θ).

Proposition 1 *Given $\theta \geq 0$ the dual problem of (P_θ) can be written equivalently as*

$$\begin{aligned} \min_{z_i, \boldsymbol{u}^i} \quad & \tfrac{2}{K^2(K+\theta)} \sum_{i=1}^{K} \|z_i - \boldsymbol{p}\|^2 \\ \text{s.t.} \quad & z_i \in \boldsymbol{E}(\boldsymbol{\mu}_i, S_i, \kappa_i), \ i = 1, \ldots, K, \\ & \boldsymbol{p} = \tfrac{1}{K} \sum_{i=1}^{K} z_i, \end{aligned} \tag{D_θ}$$

where the set $\boldsymbol{E}(\boldsymbol{\mu}, S, \kappa)$ is defined by

$$\boldsymbol{E}(\boldsymbol{\mu}, S, \kappa) = \{z \in \Re^n : z = \boldsymbol{\mu} - \kappa S \boldsymbol{u}, \ \|\boldsymbol{u}\| \leq 1\}$$

and it corresponds to an ellipsoid centered at $\boldsymbol{\mu}$ whose shape is determined by S and with its size determined by κ.

The proof of Proposition 1 is presented in the Appendix A.2.

Remark 1 It is important to mention that the objective function in problem (D_θ) is continuous and the feasible set is nonempty and compact. Therefore, by *strong duality theorem* (see [3, 28]) we have that

$$v(P_\theta) = v(D_\theta)$$

where $v(\cdot)$ denotes the optimal value of the problem.

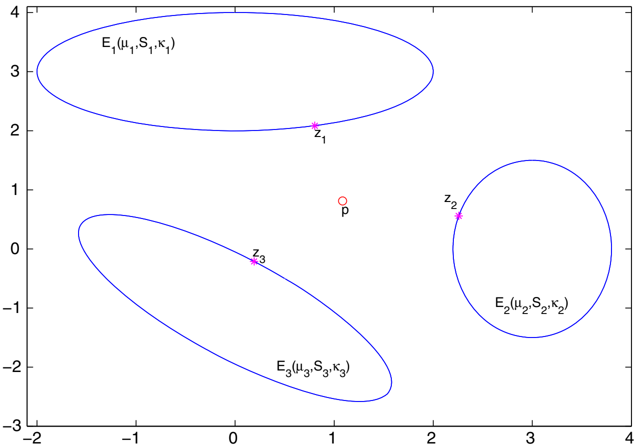

Fig. 1 Geometric illustration of problem (D_θ)

Remark 2 The obtained formulation (D_θ) is similar to that proposed by Jensen et al. [16] and Ñanculet et al. [1]. In these papers each ellipsoid is replaced by a convex hull of some training dataset, and the vector $\mathbf{p}$ is called *configuration center*. On the other hand, it is not difficult to see that, for any $\mathbf{z}_i \in \Re^n$, $i = 1, \ldots, K$,

$$\frac{1}{K} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \|\mathbf{z}_i - \mathbf{z}_j\|^2 = \sum_{i=1}^{K} \|\mathbf{z}_i - \frac{1}{K} \sum_{j=1}^{K} \mathbf{z}_j\|^2. \tag{18}$$

Then, the problem (1) can be written equivalently as:

$$\begin{aligned} \min_{\mathbf{z}_i, \mathbf{u}^i} \quad & \tfrac{2}{K^3(K+\theta)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \|\mathbf{z}_i - \mathbf{z}_j\|^2 \\ \text{s.t.} \quad & \mathbf{z}_i \in \mathbf{E}(\boldsymbol{\mu}_i, S_i, \kappa_i), \ i = 1, \ldots, K. \end{aligned} \tag{19}$$

Thus, the dual problem (D_θ) can be seen as finding the minimum distance between a set of K ellipsoids.

In Fig. 1 we illustrate the points in the ellipsoids (in 2D) obtained by using the formulation (D_θ).

The following result relates the primal and dual variables of the CC-SOCP-SVM$_l$ formulation, which is relevant since we can solve the dual formulations and then obtain the decision functions. Their proofs are presented in Appendix A.3 and A.4.

Proposition 2 *Given $\theta > 0$, we consider $\boldsymbol{w}_i \in \Re^n$ and $z_i \in \Re^n$, for $i = 1, \ldots, K$, solutions of (P_θ) and (D_θ) respectively. Then*

$$\boldsymbol{w}_i = \frac{K}{\sum_{i=1}^{K} \|z_i - \boldsymbol{p}\|^2} (z_i - \boldsymbol{p}), \quad i = 1, \ldots, K. \tag{20}$$

Remark 3 From the previous Proposition 2 it follows *a posteriori* that $\bar{\mathbf{w}} = 0$, where $\bar{\mathbf{w}}$ is defined in (10). Additionally, from (18) we have that

$$\frac{1}{K}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K}\|\mathbf{w}_i - \mathbf{w}_j\|^2 = \sum_{i=1}^{K}\|\mathbf{w}_i - \bar{\mathbf{w}}\|^2.$$

Therefore Problem (10) can be written equivalently as

$$\begin{array}{ll} \min\limits_{\mathbf{w}_i, b_i} & \frac{K+\theta}{2}\sum\limits_{i=1}^{K}\|\mathbf{w}_i\|^2 \\ \text{s.t.} & \kappa_i\|S_i^\top \mathbf{w}_i\| \leq \mathbf{w}_i^\top \boldsymbol{\mu}_i + b_i - 1,\ i = 1,\dots,K, \\ & \sum\limits_{i=1}^{K}\mathbf{w}_i = 0, \quad \sum\limits_{i=1}^{K} b_i = 0. \end{array} \tag{21}$$

Proposition 3 *Given $\theta > 0$, and $z_i \in \Re^n$, for $i = 1,\dots,K$, solution of* (1)*. Suppose that Σ_i are symmetric positive definite matrices, and $z_i \neq \boldsymbol{p}$ for all $i = 1,\dots,K$. Then, b_i solution of (D_θ) for $i = 1,\dots K$, can be written in terms of z_i as follows:*

$$b_i = 1 - \frac{K}{\sum_{i=1}^{K}\|z_i - \boldsymbol{p}}^{2}\|(z_i - \boldsymbol{p})^\top z_i, \quad i = 1,\dots,K. \tag{22}$$

In this case, the decision functions are given by

$$f^i(\boldsymbol{x}) := \frac{K}{\sum_{i=1}^{K}\|z_i - \boldsymbol{p}\|^2}(z_i - \boldsymbol{p})^\top(\boldsymbol{x} - z_i) + 1, \quad i = 1,\dots,K. \tag{23}$$

4 Kernel-based Center-of-the-Configuration SOCP-SVM formulation

Motivated by the works of [8] and [26], we modify problem (D_θ) in this section in order to include nonlinear Kernels in its definition.

Let A^i be an $n \times m_i$ matrix whose columns are the points in $\mathcal{A}^i$, and $\mathbb{X} = [A^1 A^2 \dots A^K] \in \Re^{n\times m}$ be a matrix containing the whole training set (sorted by class). Since $\mathbf{w}_i \in \Re^n$, it exists a matrix M whose columns vectors form a basis orthogonal to the span of $\mathbb{X}$. Variables $\boldsymbol{\alpha}_i$ and $\mathbf{r}_i$ are vectors of appropriate dimension such that $\mathbf{w}_i = \mathbb{X}\boldsymbol{\alpha}_i + M\mathbf{r}_i$. On the other hand, the empirical estimates of the mean and covariance are given by

$$\boldsymbol{\mu}_i = \hat{\boldsymbol{\mu}}_i = \frac{1}{m_i}A^i\mathbf{e}, \quad \Sigma_i = \hat{\Sigma}_i = S_i S_i^\top$$

with $S_i = \frac{1}{\sqrt{m_i}}(A^i - \boldsymbol{\mu}_i\mathbf{e}^\top)$

for $i = 1,\dots,K$. Then,

$$\mathbf{w}_i^\top\boldsymbol{\mu}_i = \boldsymbol{\alpha}_i^\top\mathbf{g}_i, \quad \mathbf{w}_i^\top\Sigma_i\mathbf{w}_i = \boldsymbol{\alpha}_i^\top\mathbf{G}_i\boldsymbol{\alpha}_i,$$

$$\bar{\mathbf{w}}^\top\boldsymbol{\mu}_i = \frac{1}{K}\sum_{j=1}^{K}\boldsymbol{\alpha}_j^\top\mathbf{g}_i,$$

where

$$\mathbf{g}_i = \frac{1}{m_i}\begin{bmatrix}\mathbf{K}_{1i}\mathbf{e} \\ \vdots \\ \mathbf{K}_{Ki}\mathbf{e}\end{bmatrix},$$

$$\mathbf{G}_i = \frac{1}{m_i}\begin{bmatrix}\mathbf{K}_{1i} \\ \vdots \\ \mathbf{K}_{Ki}\end{bmatrix}\left(I_{m_i} - \frac{1}{m_i}\mathbf{e}\mathbf{e}^\top\right)\left[\mathbf{K}_{1i}^\top \cdots \mathbf{K}_{Ki}^\top\right],$$

with $\mathbf{K}_{ij} = (\mathbf{K}_{ji})^\top = {A^i}^\top A^j$ matrices whose elements are inner products of data points. For instance, the entry (l,s) for the matrix $\mathbf{K}_{ij}$ is $(\mathbf{K}_{ij})_{ls} = (\mathbf{x}_l^i)^\top\mathbf{x}_s^j$. Hence, in order to obtain a Kernel formulation of Problem (21), we replace the inner product above by any function $\mathcal{K}\colon \Re^n \times \Re^n \to \Re$ satisfying the Mercer's condition [25]. Using this kernel function, the quantity $(\mathbf{x}_l^i)^\top\mathbf{x}_s^j$ is replaced by

$$(\mathbf{K}_{ij})_{ls} = \mathcal{K}(\mathbf{x}_l^i, \mathbf{x}_s^j).$$

Typical choices for this function are *the Gaussian kernel* defined by $\mathcal{K}(\mathbf{u},\mathbf{v}) = \exp(-\|\mathbf{u} - \mathbf{v}\|^2/2\sigma^2)$ with $\sigma \in \Re$ or the *polynomial* function $\mathcal{K}(\mathbf{u},\mathbf{v}) = (\mathbf{u}^\top\mathbf{v} + 1)^d$ with $d \in \mathbb{N}$ [29].

Let us denote the symmetric matrix formed with the blocks $\mathbf{K}_{ij}$ by $\mathbf{K} \in \Re^{m\times m}$. It should be noted that Mercer's condition ensures the positive semidefiniteness of the matrix $\mathbf{K}$. Then, the nonlinear formulation is given by

$$\begin{array}{ll} \min\limits_{\boldsymbol{\alpha}_i, b_i} & \frac{1}{2}\sum\limits_{i=1}^{K}\boldsymbol{\alpha}_i^\top\mathbf{K}\boldsymbol{\alpha}_i \\ \text{s.t.} & \kappa_i\sqrt{\boldsymbol{\alpha}_i^\top\mathbf{G}_i\boldsymbol{\alpha}_i} \leq \boldsymbol{\alpha}_i^\top\mathbf{g}_i + b_i - 1,\ i = 1,\dots,K, \\ & \sum\limits_{i=1}^{K}\boldsymbol{\alpha}_i = 0, \quad \sum\limits_{i=1}^{K} b_i = 0. \end{array} \tag{24}$$

Suppose that $\mathbf{K}$ is positive definite; then we can use the Cholesky factorization $\mathbf{K} = \mathbf{L}^\top\mathbf{L}$ to obtain a full rank matrix $\mathbf{L} \in \Re^{m\times m}$. Thus, introducing a new variable $\mathbf{v}_i = \mathbf{L}\boldsymbol{\alpha}_i$, for $i = 1,\dots,K$, the formulation (24) is rewritten as follows:

$$\begin{array}{ll} \min\limits_{\mathbf{v}_i, b_i} & \frac{1}{2}\sum\limits_{i=1}^{K}\|\mathbf{v}_i\|^2 \\ \text{s.t.} & \kappa_i\sqrt{\mathbf{v}_i^\top\mathbf{H}_i\mathbf{v}\ _i} \leq \mathbf{v}_i^\top\mathbf{h}_i + b_i - 1,\ i = 1,\dots,K, \\ & \sum\limits_{i=1}^{K}\mathbf{L}^{-1}\mathbf{v}_i = 0, \quad \sum\limits_{i=1}^{K} b_i = 0, \end{array} \tag{25}$$

where $\mathbf{h}_i = \mathbf{L}^{-\top}\mathbf{g}_i$, and $\mathbf{H}_i = \mathbf{L}^{-\top}\mathbf{G}_i\mathbf{L}^{-1}$, for $i = 1,\dots,K$.

Again, since $\mathbf{H}_i$ is positive semi-definite, it can be written as $\mathbf{H}_i = \mathbf{D}_i \mathbf{D}_i^\top$. Then, (25) can be written as the following quadratic second-order cone programming problem:

$$\begin{aligned} \min_{\mathbf{v}_i, b_i} \quad & \frac{1}{2} \sum_{i=1}^{K} \|\mathbf{v}_i\|^2 \\ \text{s.t.} \quad & \kappa_i \|\mathbf{D}_i^\top \mathbf{v}_i\| \leq \mathbf{v}_i^\top \mathbf{h}_i + b_i - 1, \; i = 1, \ldots, K, \\ & \sum_{i=1}^{K} \mathbf{v}_i = 0, \quad \sum_{i=1}^{K} b_i = 0. \end{aligned} \tag{26}$$

This formulation is similar to that proposed in (21), up to the factor $(K + \theta)$. Thus, proceeding similarly to the analysis of the previous section (cf. Proposition 1) we can compute the dual problem of (26), which is given by

$$\begin{aligned} \min_{\mathbf{z}_i, \mathbf{p}} \quad & \sum_{i=1}^{K} \|\mathbf{z}_i - \mathbf{p}\|^2 \\ \text{s.t.} \quad & \mathbf{z}_i \in \mathbf{E}(\mathbf{h}_i, \mathbf{D}_i, \kappa_i), \; i = 1, \ldots, K. \end{aligned} \tag{27}$$

Problem (26) will be called Kernel-based Center-of-the-Configuration SOCP-SVM formulation (CC-SOCP-SVM$_k$). According to Formulation (27), this problem can also be interpreted as a minimization of distance between ellipsoids. Finally, let us consider $\mathbb{X}_{\bullet j}$ as the jth column of the matrix $\mathbb{X}$ and, given $\mathbf{x} \in \Re^n$, we define the row vector $\mathcal{K}(\mathbf{x}, \mathbb{X})$ by

$$\mathcal{K}(\mathbf{x}, \mathbb{X}) = [\mathcal{K}(\mathbf{x}, \mathbb{X}_{\bullet 1}), \mathcal{K}(\mathbf{x}, \mathbb{X}_{\bullet 2}), \cdots, \mathcal{K}(\mathbf{x}, \mathbb{X}_{\bullet m})].$$

With this notation, we set the classification functions as

$$f_k(\mathbf{x}) = \mathcal{K}(\mathbf{x}, \mathbb{X})\boldsymbol{\alpha}_k + b_k \quad k = 1, \ldots, K.$$

5 Alternative multi-class SOCP-SVM formulations

In this section we formalize the One-versus-All and One-versus-One extensions to multiclass SOCP-SVM for both linear and kernel-based classification. These formulations were previously used in López and Maldonado [20] in the context of feature selection for microarray classification, but only as linear classifiers.

5.1 One-versus-All SOCP-SVM

The SOCP-SVM formulation for binary classification by Nath and Bhattacharyya [26] can be easily extended to OvA classification. Following the notation used in Section 3.1, the following quadratic chance-constrained programming problem is proposed for each class $k = 1, \ldots, K$:

$$\begin{aligned} \min_{\mathbf{w}_k, b_k} \quad & \frac{1}{2} \|\mathbf{w}_k\|^2 \\ \text{s.t.} \quad & \inf_{\mathbf{X}_k \sim (\boldsymbol{\mu}_k, \Sigma_k)} \Pr\{\mathbf{w}_k{}^\top \cdot \mathbf{X}_k + b_k \geq 1\} \geq \eta_k, \\ & \inf_{\mathbf{X}_k^c \sim (\boldsymbol{\mu}_k^c, \Sigma_k^c)} \Pr\{\mathbf{w}_k{}^\top \cdot \mathbf{X}_k^c + b_k \leq -1\} \geq \eta_k^c, \end{aligned} \tag{28}$$

where $\mathbf{X}_k^c$ is a random variable that generates samples of all classes but k, having $(\boldsymbol{\mu}_k^c, \Sigma_k^c)$, with $\Sigma_k, \Sigma_k^c \in \Re^{n \times n}$ symmetric positive semidefinite matrices. Again, the use of the Chebyshev-Cantelli inequality leads to the following quadratic SOCP formulation (OvA-SOCP-SVM), for each $k = 1, \ldots, K$:

$$\begin{aligned} \min_{\mathbf{w}_k, b_k, t_k} \quad & \frac{1}{2} \|\mathbf{w}_k\|^2 \\ \text{s.t.} \quad & \mathbf{w}_k^\top \boldsymbol{\mu}_k + b_k \geq 1 + \kappa_k \sqrt{\mathbf{w}_k^\top \Sigma_k \mathbf{w}_k}, \\ & -(\mathbf{w}_k^\top \boldsymbol{\mu}_k^c + b_k) \geq 1 + \kappa_k^c \sqrt{\mathbf{w}_k^\top \Sigma_k^c \mathbf{w}_k}, \end{aligned} \tag{29}$$

with $\kappa_k = \sqrt{\frac{\eta_k}{1 - \eta_k}}$ (resp. $\kappa_k^c = \sqrt{\frac{\eta_k^c}{1 - \eta_k^c}}$).

The decision rule for a new data point $\mathbf{x}$ follows: $\mathbf{x}$ belongs to the class k^* iff $k^* = \arg\max_{k=1,\ldots,K} \{\mathbf{w}_k^\top \mathbf{x} + b_k\}$.

The above formulation also can be extended to nonlinear kernel by using the same arguments of Section 4. In this case, the k-th OvA-SOCP-SVM solves the following formulation:

$$\begin{aligned} \min_{\boldsymbol{\alpha}_k, b_k} \quad & \frac{1}{2} \boldsymbol{\alpha}_k^\top \mathbf{K}^k \boldsymbol{\alpha}_k \\ \text{s.t.} \quad & \boldsymbol{\alpha}_k{}^\top \mathbf{g}_k + b_k \geq 1 + \kappa_k \sqrt{\boldsymbol{\alpha}_k{}^\top G_k \boldsymbol{\alpha}_k}, \\ & -\boldsymbol{\alpha}_k{}^\top \mathbf{g}_k^c - b_k \geq 1 + \kappa_k^c \sqrt{\boldsymbol{\alpha}_k^\top G_k^c \boldsymbol{\alpha}_k}, \end{aligned} \tag{30}$$

where $\mathbf{K}^k = [\mathbf{K}_{11}^k, \mathbf{K}_{12}^k; \mathbf{K}_{21}^k, \mathbf{K}_{22}^k] \in \Re^{m \times m}$,

$$\mathbf{g}_k = \frac{1}{m_k} \begin{bmatrix} \mathbf{K}_{11}^k \mathbf{e} \\ \mathbf{K}_{21}^k \mathbf{e} \end{bmatrix}, \; \mathbf{g}_k^c = \frac{1}{m_k^c} \begin{bmatrix} \mathbf{K}_{12}^k \mathbf{e} \\ \mathbf{K}_{22}^k \mathbf{e} \end{bmatrix},$$

$$\mathbf{G}_k = \frac{1}{m_k} \begin{bmatrix} \mathbf{K}_{11}^k \\ \mathbf{K}_{21}^k \end{bmatrix} \left(I_{m_k} - \frac{1}{m_k} \mathbf{e}\mathbf{e}^\top \right) \left[\mathbf{K}_{11}^k{}^\top \; \mathbf{K}_{21}^k{}^\top \right],$$

$$\mathbf{G}_k^c = \frac{1}{m_k^c} \begin{bmatrix} \mathbf{K}_{12}^k \\ \mathbf{K}_{22}^k \end{bmatrix} \left(I_{m_k^c} - \frac{1}{m_k^c} \mathbf{e}\mathbf{e}^\top \right) \left[\mathbf{K}_{12}^k{}^\top \; \mathbf{K}_{22}^k{}^\top \right],$$

with $\mathbf{K}_{11}^k = \mathcal{K}(A^k, A^k)$, $\mathbf{K}_{12}^k = (\mathbf{K}_{21}^k)^\top = \mathcal{K}(A^k, (A^k)^c)$, $\mathbf{K}_{22}^k = \mathcal{K}((A^k)^c, (A^k)^c)$. Here, $(A^k)^c \in \mathbb{R}^{n \times m_k^c}$ denotes a matrix whose columns are the points of all classes but k.

5.2 One-versus-One SOCP-SVM

Similar to the OvA-SOCP formulation, let $\mathbf{X}_k$ be a random variable that generates samples of class k, with mean and covariance matrix given by $(\boldsymbol{\mu}_k, \Sigma_k)$ for $k = 1, \ldots, K$, where $\Sigma_k \in \Re^{n \times n}$ are symmetric positive semidefinite matrices. Based on the idea of OvO-SVM described in Section 2.2, we can formulate an OvO version for SOCP-SVM. More precisely, for training examples from the k-th and the l-th classes ($k < l$), we solve the following quadratic chance-constrained programming problem:

$$\begin{aligned} \min_{\mathbf{w}_{kl}, b_{kl}} \quad & \frac{1}{2} \|\mathbf{w}_{kl}\|^2 \\ \text{s.t.} \quad & \inf_{\mathbf{X}_k \sim (\boldsymbol{\mu}_k, \Sigma_k)} \Pr\{\mathbf{w}_{kl}{}^\top \cdot \mathbf{X}_k + b_{kl} \geq 1\} \geq \eta_{kl}, \\ & \inf_{\mathbf{X}_l \sim (\boldsymbol{\mu}_l, \Sigma_l)} \Pr\{\mathbf{w}_{kl}{}^\top \cdot \mathbf{X}_l + b_{kl} \leq -1\} \geq \eta_{lk}, \end{aligned} \tag{31}$$

where $\eta_{kl}, \eta_{lk} \in (0, 1)$. Again, thanks to an appropriate application of the multivariate Chebyshev-Cantelli inequality, Formulation (31) can be rewritten as the following quadratic SOCP problem (OvO-SOCP):

$$\begin{aligned} \min_{\mathbf{w}_{kl},b_{kl}} \quad & \tfrac{1}{2}\|\mathbf{w}_{kl}\|^2 \\ \text{s.t.} \quad & \mathbf{w}_{kl}^\top \cdot \boldsymbol{\mu}_k + b_{kl} \geq 1 + \kappa_{kl}\sqrt{\mathbf{w}_{kl}^\top \Sigma_k \mathbf{w}_{kl}}, \\ & -\mathbf{w}_{kl}^\top \cdot \boldsymbol{\mu}_l - b_{kl} \geq 1 + \kappa_{lk}\sqrt{\mathbf{w}_{kl}^\top \Sigma_l \mathbf{w}_{kl}}, \end{aligned} \tag{32}$$

with $\kappa_{kl} = \sqrt{\frac{\eta_{kl}}{1-\eta_{kl}}}$ (resp. $\kappa_{lk} = \sqrt{\frac{\eta_{lk}}{1-\eta_{lk}}}$). Similarly to OvO-SVM, this method constructs $K(K-1)/2$ binary classifiers, one for each pair of classes. The decision function is given by $f_{kl}(\mathbf{x}) = \mathbf{w}_{kl}^\top \cdot \mathbf{x} + b_{kl}$, and the prediction of a new point $\mathbf{x}$ is done by the Max-Wins voting strategy (see Section 2.2).

Formulation (32) also can be extended to nonlinear kernel by using the same arguments of Section 4. In this case, considering training points from the k-th and the l-th classes ($k < l$), OvO-SOCP-SVM solves the following problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha}_{kl},b_{kl}} \quad & \tfrac{1}{2}\boldsymbol{\alpha}_{kl}^\top \mathbf{K}^{kl}\boldsymbol{\alpha}_{kl} \\ \text{s.t.} \quad & \boldsymbol{\alpha}_{kl}{}^\top \mathbf{g}_k + b_{kl} \geq 1 + \kappa_{kl}\sqrt{\boldsymbol{\alpha}_{kl}{}^\top G_k \boldsymbol{\alpha}_{kl}}, \\ & -\boldsymbol{\alpha}_{kl}{}^\top \mathbf{g}_l - b_{kl} \geq 1 + \kappa_{lk}\sqrt{\boldsymbol{\alpha}_{kl}^\top G_l \boldsymbol{\alpha}_{kl}}, \end{aligned} \tag{33}$$

where $\mathbf{K}^{kl} = [\mathbf{K}_{kk}, \mathbf{K}_{kl}; \mathbf{K}_{lk}, \mathbf{K}_{ll}] \in \Re^{m_k+m_l \times m_k+m_l}$,

$$\mathbf{g}_k = \frac{1}{m_k}\begin{bmatrix} \mathbf{K}_{kk}\mathbf{e} \\ \mathbf{K}_{lk}\mathbf{e} \end{bmatrix}, \quad \mathbf{g}_l = \frac{1}{m_l}\begin{bmatrix} \mathbf{K}_{kl}\mathbf{e} \\ \mathbf{K}_{ll}\mathbf{e} \end{bmatrix},$$

$$\mathbf{G}_k = \frac{1}{m_k}\begin{bmatrix} \mathbf{K}_{kk} \\ \mathbf{K}_{lk} \end{bmatrix}\left(I_{m_k} - \frac{1}{m_k}\mathbf{e}\mathbf{e}^\top\right)\begin{bmatrix} \mathbf{K}_{kk}^\top \; \mathbf{K}_{lk}^\top \end{bmatrix},$$

$$\mathbf{G}_l = \frac{1}{m_l}\begin{bmatrix} \mathbf{K}_{kl} \\ \mathbf{K}_{ll} \end{bmatrix}\left(I_{m_l} - \frac{1}{m_l}\mathbf{e}\mathbf{e}^\top\right)\begin{bmatrix} \mathbf{K}_{kl}^\top \; \mathbf{K}_{ll}^\top \end{bmatrix},$$

with $\mathbf{K}_{kk} = \mathcal{K}(A^k, A^k)$, $\mathbf{K}_{kl} = (\mathbf{K}_{lk})^\top = \mathcal{K}(A^k, A^l)$, $\mathbf{K}_{ll} = \mathcal{K}(A^l, A^l)$.

6 Experimental results

We applied the proposed SOCP-SVM approach in its linear and kernel-based form to five well-known benchmark datasets for multi-class classification. We also used other alternative multi-class SVM formulations described in Section 2 (MC-SVM, OvO-SVM, OvA-SVM, and AD-SVM) for comparison purposes.

This section is organized as follows. We provide a description of the datasets in Section 6.1, while Section 6.2 presents a summary of the performance obtained for all the approaches and a detailed discussion of these results.

6.1 Datasets and experimental settings

For our experiments we used five datasets available from the UCI Machine Learning Repository [5]: Iris, Wine, Glass, Waveform, and Segment. We also include a sixth dataset used in a previous research project for classification of fish schools (see [8] for more details). Table 1 summarizes the descriptive information for each dataset.

Table 1 Number of examples, number of variables and number of classes for all datasets

Dataset	#examples	#variables	#classes
Iris	150	4	3
Wine	178	13	3
Glass	214	13	6
Fish	762	12	3
Segment	2310	19	7
Waveform	5000	21	3

The following model selection procedure was performed: 10-fold cross-validation was used for model selection, where balanced accuracy is monitored to assess predictive performance. This measure is computed as follows: the recall of each class is first obtained, and then averaged over the number of different classes. We studied the following values of $\eta_k \in \{0.2, 0.4, 0.6, 0.8\}$. For standard SVM approaches, we used the following set of values for parameters C and σ (only for kernel methods): $\{2^{-7}, 2^{-6}, \ldots, 2^{7}\}$. These exponentially growing sequences for (C, σ) are recommended in [15, 22]. We used the Spider Toolbox for Matlab [34] for standard SVM approaches, and the SeDuMi Matlab Toolbox for SOCP-based classifiers [30].

6.2 Classification performance summary

Table 2 summarizes the results for all linear approaches. The best performance among all methods in terms of balanced accuracy is highlighted in bold type.

It can be seen in Table 2 that no method outperformed others, although the SOCP-SVM methods achieved best results on four out of six datasets and, in particular, CC-SOCP-SVM had best predictive performance in Waveform dataset. From the traditional approaches, OVO-SVM achieved best results on two out of six datasets. Notice that CC-SOCP-SVM has always better performance than AD-SVM, demonstrating the advantage of using ellipsoids

Table 2 Performance summary for all linear classification approaches

	Iris	Wine	Glass	Fish	Segment	Waveform
MC-SVM$_l$	96.0	99.0	57.3	69.7	90.9	87.3
OVA-SVM$_l$	94.7	98.6	60.7	74.4	92.7	87.0
OVO-SVM$_l$	**98.0**	98.6	66.1	**80.0**	95.6	87.0
AD-SVM$_l$	95.3	93.2	51.0	67.8	80.1	87.0
CC-SOCP-SVM$_l$	96.0	98.5	60.6	68.0	85.5	**87.4**
OVA-SOCP-SVM$_l$	96.7	**99.1**	**68.9**	67.3	91.0	86.7
OVO-SOCP-SVM$_l$	97.3	**99.1**	**68.9**	77.2	**97.0**	87.1

Table 3 Performance summary for all kernel-based classification approaches

	Iris	Wine	Glass	Fish	Segment	Waveform
MC-SVM$_k$	97.3	99.0	71.4	83.2	**98.3**	87.0
OVA-SVM$_k$	97.3	**99.5**	71.8	81.6	97.5	87.2
OVO-SVM$_k$	98.0	99.0	72.2	82.6	97.4	87.0
AD-SVM$_k$	96.7	97.6	61.2	74.5	93.8	87.0
CC-SOCP-SVM$_k$	96.7	98.9	61.9	77.1	86.8	**87.8**
OVA-SOCP-SVM$_k$	97.3	99.1	75.0	84.4	96.5	87.3
OVO-SOCP-SVM$_k$	**98.7**	**99.5**	**76.3**	**87.1**	97.6	87.3

instead of reduced convex hulls to compute the center of the configuration.

Table 3 summarizes the results for all kernel-based approaches. Again, the best performance among kernel methods in terms of balanced accuracy is highlighted in bold type.

Similarly to the linear case, in Table 3 we observe superior performance for the SOCP-SVM methods based on kernel functions, achieving best results on five out of six datasets. Again, our CC-SOCP-SVM approach performed better on the Waveform dataset, while the OVO-SOCP-SVM method behaved better for the first four datasets. The proposed CC-SOCP-SVM method performs better than AD-SVM, being worse than the latter method only in one dataset (Segment). In general, the robust counterparts of the explored SVM methods OVA, OVO, and AD-SVM have better overall results than the original approaches.

From previous experiments we conclude that:

- Kernel-based versions perform better in general, and should be considered for multi-class classification. Our approach has the advantage that it can be extended to kernel methods, compared with the linear methods that have been suggested in the literature.
- Best overall performance is achieved with OVO-SOCP-SVM, although no method outperformed others. Our proposals achieve best overall results, and are recommended as alternatives for SVM classification.
- Our CC-SOCP-SVM method performs better than AD-SVM, which follows a similar geometric principle (the center of the configuration), demonstrating the usefulness of the robust optimization scheme based on second-order cones.

7 Conclusions

In this work, we present a novel multi-class SVM approach based on the principle of the center of the configuration [1, 16] and second-order cone programming [26]. The concept of center of the configuration is an appealing geometric principle: it corresponds to the point from which all classes are equidistant, while the use of second-order cones confers robustness to the proposal, given their ability to generalize the training patterns better by assuming the worst distribution of the data. We identified the following advantages of the CC-SOCP-SVM method according to the results of our experiments presented in the previous section:

- It can be extended to kernel methods, conferring flexibility to the classifier and improving predictive results.
- It solves an SOCP problem based on a balanced design, in which each conic constraint corresponds to a particular class pattern that should be correctly classified up to the rate η. This approach yields to the benefit of the correct generalization of all classes compared to standard SVM, where each constraint is related to a training sample, biasing the classification to the majority class when facing class-imbalance and overlap [23].
- It solves a single optimization problem, constructing all classifiers simultaneously by taking all available information into account. This strategy differs from OvO and OvA approaches, in which the classification functions are obtained by solving independent problems. Additionally, the OvO-SVM and MC-SVM methods require the construction of several classifiers, one for each pair of classes, and therefore the running times and complexity grow exponentially with the number of classes. Our strategy constructs fewer classifiers and therefore is capable of solving problems that have a high number of classes.
- It has better predictive performance than alternative multi-class approaches based on the principle of the center of the configuration, such as AD-SVM, demonstrating the effectiveness of the robust strategy based on second-order cones.

In this work we also extend the One-versus-All and One-versus-One strategies to SOCP-SVM with excellent results in both linear and kernel-based versions. In fact, the OvO-SOCP-SVM method has the best overall results among all the methods studied. Since no method outperformed all the others, we recommend our approach as part of a pool of methods for multi-class classification. Since our proposals follow different classification strategies compared with traditional SVM methods, they are also good candidates for the construction of ensembles for SVM classification [23].

Some research opportunities for future work have been identified. The optimization process for second-order cone programming formulations is, in general, more time consuming than quadratic programming methods, such as traditional SVM. Although several techniques have been suggested for an efficient optimization process [2, 19], none of these strategies has been adapted for SOCP-based SVM. Faster SOCP-SVM implementations that exploit the structure of the problem are needed in order to be able to apply such techniques on large scale datasets. Additionally, the proposed method has interesting properties for class-imbalanced classification, due to its balanced design. The use of this formulation in multi-class applications with skewed class distributions [13] is suggested.

Acknowledgments The first author was supported by FONDECYT project 11110188 and by CONICYT Anillo ACT1106, the second one was funded by FONDECYT project 11121196, and third author was supported by FONDECYT project 1130905. The work reported in this paper has been partially funded by the Complex Engineering Systems Institute (ICM: P-05-004-F, CONICYT: FB016).

Appendix A: Dual formulation for multi-class SOCP-SVM

A.1 Proof of Theorem 1

Proof The Lagrangian function associated with Problem (P_θ) is given by:

$$\begin{aligned} L(\tilde{\mathbf{w}}, \mathbf{b}, \alpha_i, \beta) &= \frac{1}{2}\|\mathbf{Q}^{1/2}(\theta)\tilde{\mathbf{w}}\|^2 + \sum_{i=1}^{K} \alpha_i (\kappa_i \| S_i^\top H^i \tilde{\mathbf{w}}\| \\ &\quad -(H^i\tilde{\mathbf{w}})^\top \boldsymbol{\mu}_i - (\mathbf{d}^i)^\top \mathbf{b}) \\ &\quad + \sum_{i=1}^{K} \alpha_i + \beta(\mathbf{e}^\top \mathbf{b}). \end{aligned}$$

Since the relationship $\|\mathbf{v}\| = \max_{\|\mathbf{u}\|\le 1} \mathbf{v}^\top \mathbf{u}$ holds for any $\mathbf{v} \in \Re^n$, we can rewrite the Lagrangian as follows

$$\begin{aligned} L(\tilde{\mathbf{w}}, \mathbf{b}, \alpha_i, \beta) &= \max_{\mathbf{u}^i}\{L_1(\tilde{\mathbf{w}}, \mathbf{b}, \alpha_i, \beta, \mathbf{u}^i) : \|\mathbf{u}^i\| \\ &\qquad \le 1, i = 1, \dots, K\}, \end{aligned}$$

where L_1 is given by

$$\begin{aligned} &L_1(\tilde{\mathbf{w}}, \mathbf{b}, \alpha_i, \beta, \mathbf{u}^i) \\ &= \frac{1}{2}\|\mathbf{Q}^{1/2}(\theta)\tilde{\mathbf{w}}\|^2 \\ &\quad + \sum_{i=1}^{K} \alpha_i \left(\kappa_i (S_i^\top H^i \tilde{\mathbf{w}})^\top \mathbf{u}^i - (H^i\tilde{\mathbf{w}})^\top \boldsymbol{\mu}_i\right) \\ &\quad + \sum_{i=1}^{K} \alpha_i (1 - (\mathbf{d}^i)^\top \mathbf{b}) + \beta(\mathbf{e}^\top \mathbf{b}). \end{aligned} \tag{34}$$

Thus, Problem (P_θ) can be written equivalently as

$$\min_{\tilde{\mathbf{w}},\mathbf{b}} \max_{\alpha_i,\beta,\mathbf{u}^i} \{L_1(\tilde{\mathbf{w}}, \mathbf{b}, \alpha_i, \beta, \mathbf{u}^i) : \|\mathbf{u}^i\| \le 1, \alpha_i \ge 0, i = 1, \dots, K\}.$$

Hence, the dual problem (see e.g. [7, 24, 28]) of (P_θ) is given by

$$\begin{aligned} &\max_{\alpha_i,\beta,\mathbf{u}^i} \min_{\tilde{\mathbf{w}},\mathbf{b}} \{L_1(\tilde{\mathbf{w}}, \mathbf{b}, \alpha_i, \beta, \mathbf{u}^i) : \|\mathbf{u}^i\| \le 1,\ \alpha_i \\ &\ge 0,\ i = 1, \dots, K\}. \end{aligned} \tag{35}$$

The expression (35) now enables us to eliminate the primal variables to give the dual. Computing the first order optimization condition of the internal minimum problem (35) yields

$$\begin{aligned} \nabla_{\tilde{\mathbf{w}}} L_1 &= \mathbf{Q}(\theta)\tilde{\mathbf{w}} + \sum_{i=1}^{K} \alpha_i \left(\kappa_i {H^i}^\top S_i \mathbf{u}^i - {H^i}^\top \boldsymbol{\mu}_i\right) \\ &= 0, \end{aligned} \tag{36}$$

$$\nabla_{\mathbf{b}} L_1 = -\sum_{i=1}^{K} \alpha_i \mathbf{d}^i + \beta \mathbf{e} = 0. \tag{37}$$

It follows from the previous (37) that $\alpha_i = \beta$ for all $i = 1, \dots, K$, and therefore, replacing in (34) we get

$$\begin{aligned} L_1(\tilde{\mathbf{w}}, \mathbf{b}, \alpha_i, \beta, \mathbf{u}^i) &= \frac{1}{2}\|\mathbf{Q}^{1/2}(\theta)\tilde{\mathbf{w}}\|^2 \\ &\quad + \beta \sum_{i=1}^{K} \tilde{\mathbf{w}}^\top H^{i\top}(\kappa_i S_i^\top \mathbf{u}^i - \boldsymbol{\mu}_i) \\ &\quad + K\beta. \end{aligned} \tag{38}$$

Additionally, from (36) we get

$$\mathbf{Q}(\theta)\tilde{\mathbf{w}} = \beta \sum_{i=1}^{K} {H^i}^\top (\boldsymbol{\mu}_i - \kappa_i S_i \mathbf{u}^i). \tag{39}$$

Replacing (39) in (38) we obtain the reduced form

$$L_1(\tilde{\mathbf{w}}, \mathbf{b}, \alpha_i, \beta, \mathbf{u}^i) = -\frac{1}{2}\|\mathbf{Q}^{1/2}(\theta)\tilde{\mathbf{w}}\|^2 + K\beta. \tag{40}$$

In order to use (39) to compute $\mathbf{Q}^{1/2}(\theta)\tilde{\mathbf{w}}$, and then to obtain the dual problem, we distinguish two main cases: $\theta > 0$, in which we have strong convexity of the objective function and we obtain good mathematical properties as uniqueness of the optimal solution; and the case $\theta = 0$, in which most of this mathematical structure is lost but we can still derive a dual problem for this alternative by using some properties of the matrix $\mathbf{Q}(0)$.

Case $\theta > 0$: In this case we note first that (see [35, Proposition 3.4] for details)

$$\mathbf{Q}^{-1/2}(\theta) = \frac{1}{\sqrt{K+\theta}} I_{nK} + \frac{\sqrt{K+\theta} - \sqrt{\theta}}{K\sqrt{\theta}\sqrt{K+\theta}} \mathcal{J},$$

then

$$\mathbf{Q}^{-1/2}(\theta){H^i}^\top = \frac{1}{\sqrt{K+\theta}} {H^i}^\top. \tag{41}$$

Therefore, multiplying (39) by $\mathbf{Q}^{-1/2}(\theta)$ we obtain

$$\mathbf{Q}^{1/2}(\theta)\tilde{\mathbf{w}} = \frac{\beta}{\sqrt{K+\theta}} \sum_{i=1}^{K} {H^i}^\top \mathbf{z}_i, \text{ where } \mathbf{z}_i = \boldsymbol{\mu}_i - \kappa_i S_i \mathbf{u}^i.$$

Replacing this expression in the reduced Lagrangian function (40), we get

$$L_1 = K\beta - \frac{\beta^2}{2(K+\theta)} \| \sum_{i=1}^{K} {H^i}^\top \mathbf{z}_i \|^2,$$

and therefore we obtain the following dual formulation

$$\begin{aligned} \max_{\mathbf{u}^i,\beta} \quad & K\beta - \tfrac{\beta^2}{2(K+\theta)} \| \sum_{i=1}^{K} {H^i}^\top \mathbf{z}_i \|^2 \\ \text{s.t.} \quad & \mathbf{z}_i = \boldsymbol{\mu}_i - \kappa_i S_i \mathbf{u}^i, \\ & \|\mathbf{u}^i\| \leq 1, \ i = 1, \ldots, K, \\ & \beta \geq 0. \end{aligned} \tag{42}$$

Note that, after some algebraic manipulation, we have

$$\| \sum_{i=1}^{K} {H^i}^\top \mathbf{z}_i \|^2 = \sum_{i=1}^{K} \|\mathbf{z}_i - \mathbf{p}\|^2, \quad \text{with} \quad \mathbf{p} = \frac{1}{K} \sum_{i=1}^{K} \mathbf{z}_i. \tag{43}$$

In consequence, Problem (42) can be written as

$$\begin{aligned} \max_{\mathbf{z}_i,\mathbf{u}^i,\beta} \quad & K\beta - \tfrac{\beta^2}{2(K+\theta)} \sum_{i=1}^{K} \|\mathbf{z}_i - \mathbf{p}\|^2 \\ \text{s.t.} \quad & \mathbf{z}_i = \boldsymbol{\mu}_i - \kappa_i S_i \mathbf{u}^i, \\ & \|\mathbf{u}^i\| \leq 1, \ i = 1, \ldots, K, \\ & \mathbf{p} = \tfrac{1}{K} \sum_{i=1}^{K} \mathbf{z}_i, \\ & \beta \geq 0. \end{aligned} \tag{44}$$

Case $\theta = 0$: We note the following property $\mathbf{Q}(0) = \sqrt{K}\mathbf{Q}^{1/2}(0)$. Replacing this in (39) one has

$$\mathbf{Q}^{1/2}(0)\tilde{\mathbf{w}} = \frac{\beta}{\sqrt{K}} \sum_{i=1}^{K} {H^i}^\top (\boldsymbol{\mu}_i - \kappa_i S_i \mathbf{u}^i).$$

Using this in (40) and proceeding in a similar way to the case $\theta > 0$, we obtain the same formulation in (44) with $\theta = 0$. □

A.2 Proof of Proposition 1

Proof It is suffices to prove that the formulation (D_θ) is equivalent to (17). We note that, given $\theta \geq 0$, the objective function of the dual problem (17) is concave with respect to β therefore attains its maximum value at

$$\beta = \frac{K(K+\theta)}{\sum_{i=1}^{K} \|\mathbf{z}_i - \mathbf{p}\|^2}, \tag{45}$$

also its optimal value is given by

$$\frac{K^2(K+\theta)}{2\sum_{i=1}^{K} \|\mathbf{z}_i - \mathbf{p}\|^2}.$$

Hence, replacing this in (17) and using the definition of $\mathbf{E}(\boldsymbol{\mu}, S, \kappa)$ we get that (17) is equivalent to problem:

$$\begin{aligned} \max_{\mathbf{z}_i,\mathbf{u}^i} \quad & \tfrac{K^2(K+\theta)}{2\sum_{i=1}^{K} \|\mathbf{z}_i - \mathbf{p}\|^2} \\ \text{s.t.} \quad & \mathbf{z}_i \in \mathbf{E}(\boldsymbol{\mu}_i, S_i, \kappa_i), \quad i = 1, \ldots, K, \\ & \mathbf{p} = \tfrac{1}{K} \sum_{i=1}^{K} \mathbf{z}_i, \end{aligned} \tag{46}$$

turn the max term to min we obtain the result. □

A.3 Proof of Proposition 2

Proof It is not difficult to see that

$$\sum_{i=1}^{K} {H^i}^\top \mathbf{z}_i = \frac{1}{K}\mathbf{Q}(0)\tilde{\mathbf{z}}, \tag{47}$$

where we denote $\tilde{\mathbf{z}} = [\mathbf{z}_1^\top, \mathbf{z}_2^\top, \ldots, \mathbf{z}_K^\top]^\top \in \Re^{nK}$. Therefore, from (39) the relation between $\mathbf{w}_i$ and $\mathbf{z}_i$ can be written equivalently as

$$\mathbf{Q}(\theta)\tilde{\mathbf{w}} = \frac{\beta}{K}\mathbf{Q}(0)\tilde{\mathbf{z}}.$$

Since $\mathbf{Q}(\theta)$ is invertible, when $\theta > 0$, we have

$$\tilde{\mathbf{w}} = \frac{\beta}{K}\mathbf{Q}(\theta)^{-1}\mathbf{Q}(0)\tilde{\mathbf{z}} = \frac{\beta}{K(K+\theta)}\mathbf{Q}(0)\tilde{\mathbf{z}}.$$

where the last inequality follows from (41) and (47). Then, we conclude that $\mathbf{w}_i$ can be written in terms of $\mathbf{z}_i$ and $\mathbf{p}$ as

$$\mathbf{w}_i = \frac{\beta}{(K+\theta)}(\mathbf{z}_i - \mathbf{p}) = \frac{K}{\sum_{i=1}^{K} \|\mathbf{z}_i - \mathbf{p}\|^2}(\mathbf{z}_i - \mathbf{p}),$$

$i = 1, \ldots, K$.

□

We note that in the case $\theta = 0$, $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{z}}$, are related by

$$\mathbf{Q}(0)\tilde{\mathbf{w}} = \frac{\beta}{K}\mathbf{Q}(0)\tilde{\mathbf{z}},$$

but the uniqueness of solution is lost, because the matrix $\mathbf{Q}(0)$ is symmetric positive semi-definite.

A.4 Proof of Proposition 3

Proof The KKT conditions of the dual formulation (1) can be summarized as follows

$$\begin{aligned} -\frac{4\kappa_i}{K^2(K+\theta)} S_i^\top(\mathbf{z}_i - \mathbf{p}) + \gamma_i \mathbf{u}_i &= 0, \\ \gamma_i(\|\mathbf{u}_i\| - 1) &= 0, \qquad (48) \\ \sum_i (\mathbf{z}_i - \mathbf{p}) &= 0, \\ \|\mathbf{u}_i\| \le 1,\ \gamma_i &\ge 0. \qquad (49) \end{aligned}$$

Since Σ_i are symmetric positive definite matrices, and $\mathbf{z}_i \neq \mathbf{p}$ for all $i = 1, \ldots, K$, we have that $\mathbf{z}_i - \mathbf{p}$ does not belong to the null space of $S_i^\top$. Consequently, the Lagrangian multipliers γ_i are strictly positive. This implies that $\|\mathbf{u}_i\| = 1$ holds. In such situation, $\mathbf{z}_i = \boldsymbol{\mu}_i - \kappa_i S_i \mathbf{u}_i$ belong to the boundary of the Ellipsoid $\mathbf{E}(\boldsymbol{\mu}_i, S_i, \kappa_i)$.

Furthermore, by (37) and (45) we have that $\alpha_i = \beta > 0$ for $i = 1, \ldots, K$. This implies that the constraints in (P_θ) are active. Then, since $\bar{\mathbf{w}} = 0$ (cf. Remark 3) and $\|\mathbf{u}_i\| = 1$, we get from (10) that

$$\kappa_i \|u_i\| \|S_i^\top \mathbf{w}_i\| = \mathbf{w}_i^\top \boldsymbol{\mu}_i + b_i - 1 \quad i = 1, \ldots, K.$$

We note at optimality $S_i^\top(\mathbf{z}_i - \mathbf{p})$ is parallel to $\mathbf{u}_i$, for $i = 1, \ldots, K$, thus, we have that $\|\mathbf{u}_i\| \|S_i^\top \mathbf{w}_i\| = \mathbf{u}_i^\top S_i^\top \mathbf{w}_i$, obtaining from the above expression that

$$0 = \mathbf{w}_i^\top (\boldsymbol{\mu}_i - \kappa_i S_i \mathbf{u}_i) + b_i - 1 \quad i = 1, \ldots, K.$$

Then, we get the following conditions:

$$\mathbf{w}_i^\top \mathbf{z}_i + b_i = 1, \qquad \text{for } i = 1, \ldots, K.$$

This geometrically means that the hyperplanes $\mathbf{w}_i^\top \mathbf{x} + b_i = 1$ are tangents to the ellipsoids $\mathbf{E}(\boldsymbol{\mu}_i, S_i, \kappa_i)$, for $i = 1, \ldots, K$. Using the above relation and (45), one can compute the value of b_i obtaining the result in (22).

Finally, as decision functions are given by $f^i(\mathbf{x}) := \mathbf{w}_i^\top \mathbf{x} + b_i$, the result in (23) is obtained by replacing the values of b_i from (22) and $\mathbf{w}_i$ from Proposition 2. □

References

1. Ñanculef R, Concha C, Allende H, Candel D, Moraga C (2009) Ad-svms: A light extension of svms for multicategory classification. Int J Hybrid Intell Syst 6(2):69–79
2. Ñanculef R, Frandi E, Sartori C, Allende H (2014) A novel frank-wolfe algorithm. analysis and applications to large-scale svm training. Inf Sci 285:66–99
3. Alizadeh F, Goldfarb D (2003) Second-order cone programming. Math Program 95:3–51
4. Alvarez F, López J, Ramírez CH (2010) Interior proximal algorithm with variable metric for second-order cone programming: applications to structural optimization and support vector machines. Optim Methods Softw 25(6):859–881
5. Asuncion A, Newman D (2007) UCI machine learning repository. http://archive.ics.uci.edu/ml/
6. Attouch H (1996) Viscosity solutions of minimization problems. SIAM J Optim 6(3):769–806
7. Bertsekas D (1982) Constrained optimization and lagrange multiplier methods. Academic Press, New York
8. Bosch P, López J, Ramírez H, Robotham H (2013) Support vector machine under uncertainty: An application for hydroacoustic classification of fish-schools in chile. Expert Syst Appl 40(10):4029–4034
9. Bravo C, Thomas L, Weber R (2015) Improving credit scoring by differentiating defaulter behaviour. J Oper Res Soc 66:771–781
10. Bredensteiner EJ, Bennett KP (1999) Multicategory classification by support vector machines. Comput Optim Appl 12:53–79
11. Debnath R, Muramatsu M, Takahashi H (2005) An efficient support vector machine learning method with second-order cone programming for large-scale problems. Appl Intell 23:219–239
12. Friedman J (1996) Another approach to polychotomous classification. Tech rep, Department of Statistics, Stanford University. http://www-stat.stanford.edu/~jhf/ftp/poly.ps.Z
13. Hao PY, Chiang JH, Lin YH (2009) A new maximal-margin spherical-structured multi-class support vector machine. Appl Intell 30:98–111
14. Hsu C, Lin C (2002) A comparison of methods for multiclass support vector machines. IEEE Trans Neural Netw 13(2):415–425
15. Hsu CW, Chang CC, Lin CJ (2010) A practical guide to support vector classification. Tech rep, Department of Computer Science, National Taiwan University
16. Jenssen R, Kloft M, Zien A, Sonnenburg S, Müller K (2012) A scatter-based prototype framework and multi-class extension of support vector machines. PLoS ONE 7(10):e42,947
17. Kressel UG (1999) Advances in kernel methods. MIT Press, Cambridge, pp 255–268. USA, chap Pairwise classification and support vector machines
18. Lanckriet G, Ghaoui L, Bhattacharyya C, Jordan M (2003) A robust minimax approach to classification. J Mach Learn Res 3:555–582
19. Li C, Liu K, Wang H (2011) The incremental learning algorithm with support vector machine based on hyperplane-distance. Appl Intell 34:19–27
20. López J, Maldonado S (2015) Feature selection for multiclass support vector machines using second-order cone programming. Intelligent Data Analysis Accepted
21. Maldonado S, López J (2014a) Alternative second-order cone programming formulations for support vector classification. Inf Sci 268:328–341
22. Maldonado S, López J (2014b) Imbalanced data classification using second-order cone programming support vector machines. Pattern Recogn 47:2070–2079
23. Maldonado S, Montecinos C (2014) Robust classification of imbalanced data using ensembles of one-class and two-class svms. Intelligent Data Analysis, Special Issue on Business Analytics and Intelligent Optimization 18:95–112
24. Mangasarian OL (1994) Nonlinear programming. Classics in Applied Mathematics, Society for Industrial and Applied Mathematics
25. Mercer J (1909) Functions of positive and negative type, and their connection with the theory of integral equations. Philos Trans R Soc Lond 209:415–446
26. Nath S, Bhattacharyya C (2007) Maximum margin classifiers with specified false positive and false negative error rates. In: Proceedings of the SIAM International Conference on Data mining
27. Rifkin R, Klautau A (2004) In defense of one-vs-all classification. J Mach Learn Res 5:101–141

28. Rockafellar RT (1997) Convex analysis. Princeton Landmarks in Mathematics. Princeton University Press, Princeton. reprint of the 1970 original, Princeton Paperbacks
29. Schölkopf B, Smola AJ (2002) Learning with Kernels. MIT Press
30. Sturm J (1999) Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. Optim Methods Softw 11(12):625–653. special issue on Interior Point Methods (CD supplement with software)
31. Tikhonov A, Arsénine V (1976) Méthodes de resolution de problèmes mal posés. Éditions Mir, Moscow. french translation of the 1974 Russian edition
32. Vapnik V (1998) Statistical learning theory. John Wiley and Sons
33. Weston J, Watkins C (1999) Multi-class support vector machines. In: Proceedings of the Seventh European Symposium on Artificial Neural Networks
34. Weston J, Elisseeff A, BakIr G, Sinz F (2005) The spider machine learning toolbox. http://www.kyb.tuebingen.mpg.de/bs/people/spider/
35. Yajima Y (2005) Linear programming approaches for multicategory support vector machines. Eur J Oper Res 162(2):514–531
36. Yang K, Cai Z, Li J, Lin G (2006) A stable gene selection in microarray data analysis. BMC Bioinformatics 7:228
37. Zhong P, Fukushima M (2007) Second-order cone programming formulations for robust multiclass classification. Neural Comput 19:258–282

Julio López received his B.S. degree in Mathematics in 2000 from the University of Trujillo, Perú. He also received the M.S. degree in Sciences in 2003 from the University of Trujillo, Perú and the Ph.D. degree in Engineering Sciences, minor Mathematical Modelling in 2009 from the University of Chile. Currently, he is an assistant Professor of Institute of Basic Sciences at the University Diego Portales, Santiago, Chile. His research interests include conic programming, convex analysis, algorithms and machine learning.

Sebastián Maldonado received his B.S. and M.S. degree from the University of Chile, in 2007, and his Ph.D. degree from the University of Chile, in 2011. He is currently Professor at the School of Engineering and Applied Sciences, Universidad de los Andes, Santiago, Chile. His research interests include statistical learning, data mining and business analytics.

Miguel Carrasco received his B.S. degree in Mathematics in 2002 and the B.S. degree in Computing Sciences in 2005 from the University of Chile. He also received the Ph.D. degree in Engineering Sciences, minor Mathematical Modelling in 2007 from the University of Chile in collaboration with University of Montpellier II, France. Currently, he is full time professor at the School of Engineering and Applied Sciences, Universidad de los Andes, Santiago, Chile. His research interests include convex analysis, proximal type algorithms, conic programming and topology optimization.